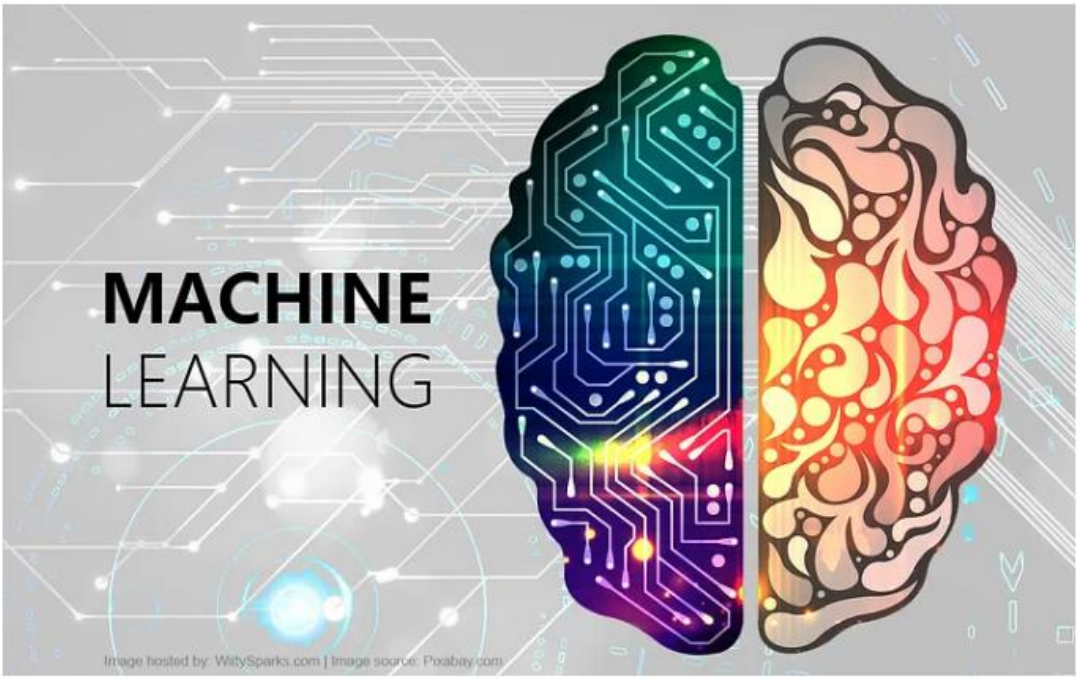
פרויקט למידת מכונה - חלק א

מגיש: נועם גנים



הגדרת הבעיה:

1. תיאור כללי של עולם התוכן הנחקר

הבעיה המחקרית בנתוני סקר שביעות רצון נוסעים של חברות תעופה היא זיהוי הגורמים המשפיעים באופן משמעותי על שביעות רצון הנוסעים, כך ששביעות רצון נוסעים היא גורם חשוב המשפיע על הצלחתן של חברות תעופה. הבנת גורמים אלו תאפשר לחברות תעופה לשפר את השירותים שלהן ,להגביר את שביעות רצון הלקוחות ובכך להגדיל את רווחיהן.

מחקרים קודמים בנושא התמקדו בזיהוי גורמים שונים המשפיעים על שביעות רצון נוסעים . גורמים אלו כוללים איכות השירות, מחיר הכרטיס, זמן הטיסה, ביטולים ועיכובים והתנהלות חברת התעופה בעת תקלות. מחקרים אלו השתמשו במגוון שיטות מחקר ,כגון ניתוח סטטיסטי של נתוני סקרי שביעות רצון , סימולציות, ראיונות עם נוסעים ותצפיות בהתנהגות נוסעים.

2. הגדרת שאלת המחקר

מהם הגורמים המשפיעים באופן משמעותי על שביעות רצון נוסעי חברות תעופה?

אנו מצפים שהכלים והשיטות של מערכות לומדות יעזרו לנו לזהות את הגורמים המשפיעים באופן משמעותי על שביעות רצון נוסעי חברות תעופה, כמו פילוח נוסעים, גילוי דפוסים. בנוסף נרצה לפתח מודלים של למידת מכונה שיאפשרו לנו לחזות את שביעות רצון הנוסעים בדיוק גבוה. אנו מצפים שהתוצאות של המחקר יהיו שימושיות לחברות תעופה כדי לשפר את רמת השירותים שלהן, להגביר את שביעות רצון הלקוחות ולשפר את התחרותיות שלהן בשוק.

הבנת הנתונים:

1. תיעוד מקורות הנתונים ומשמעותם

מסבירים	משמעות משתנה	סוג משתנה	מקור הנתון
Gender	מגדר – זכר/נקבה	קטגוריאלי לא ניתן לסידור	דרך שאלונים
Customer	סוג לקוח (לקוח נאמן/לקוח לא נאמן)	קטגוריאלי לא ניתן לסידור	לפי נתוני עבר של קריטריונים מוגדרים
Age	גיל הלקוח	רציף	דרך שאלונים

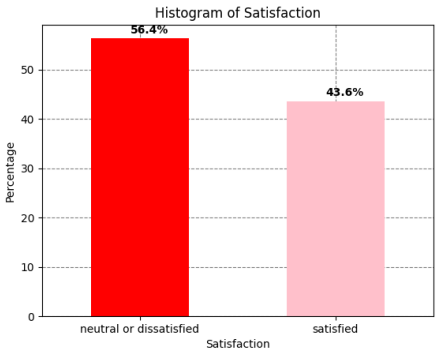
Type of Travel	סיבת הנסיעה (נסיעות אישיות/נסיעות עסקים)	קטגוריאלי לא ניתן לסידור	דרך שאלונים
Class	מחלקה לנסיעות (אקו, אקו פלוס, עסקים)	קטגוריאלי ניתן לסידור	דרך שאלונים
Flight Distance	מרחק הטיסה ב-ק"מ	רציף	מדידות
Plane colors	מספר צבעי המטוס	קטגוריאלי ניתן לסידור	ידנית
Inflight Wi-Fi service	דירוג שירות Wi-Fi בטיסה (0 עד 5)	בדיד	דרך שאלונים
Departure/Arrival time convenient	דירוג נוחות זמני יציאה/הגעה (0 עד 5)	בדיד	דרך שאלונים
Ease of Online booking	דירוג קלות ההזמנה באינטרנט (0 עד 5)	בדיד	דרך שאלונים
Gate location	דירוג מיקום השער (0 עד 5)	בדיד	דרך שאלונים
Food and drink	דירוג שירות האוכל והשתייה (0 עד 5)	בדיד	דרך שאלונים
Seat comfort	דירוג נוחות המושב (1 עד 5)	בדיד	דרך שאלונים
On-board service	דירוג השירות על הסיפון (1 עד 5)	בדיד	דרך שאלונים
Leg room service	דירוג מקום לרגליים (1 עד 5)	בדיד	דרך שאלונים
Baggage handling	דירוג טיפול בכבודה (1 עד 5)	בדיד	דרך שאלונים
Check-in service	דירוג שירות הצ'ק-אין (1 עד 5)	בדיד	דרך שאלונים
Inflight service	דירוג שירות בטיסה (1 עד 5)	בדיד	דרך שאלונים
Cleanliness	דירוג הניקיון (1 עד 5)	בדיד	דרך שאלונים
Departure Delay in Minutes	עיכוב יציאה בדקות	רציף	נתונים סטטיסטים
Arrival Delay in Minutes	עיכוב הגעה בדקות	רציף	נתונים סטטיסטים
Satisfaction משתנה המוסבר	**רמת שביעות רצון (מרוצה/ניטרלי או לא מרוצה)**	**קטגוריאלי ניתן לסידור**	**דרך שאלונים**

2. הסתברויות אפריוריות וקשרים בין מאפיינים

2.1, 2.2

משתנה מטרה:

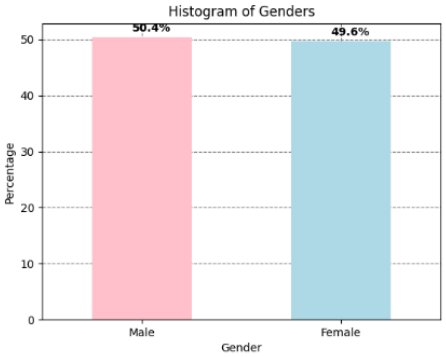
satisfaction - 0



ההסתברויות האפריוריות מראות כי 56.4% מהאנשים אינם מרוצים, ו-43.6% מרוצים. הנתונים יחסית מאוזנים ניתן לראות שמתפלגים בערך חצי חצי עם הטיה ללא מרוצים . לדעתנו הדבר מייצג את המציאות כיוון שתחושת חוסר שביעות רצון היא נפוצה בעולם.

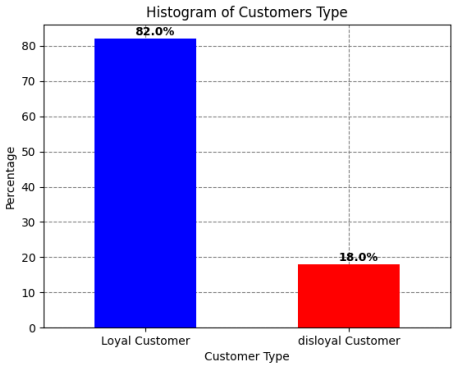
משתנים קטגוריאליים:

gender -1

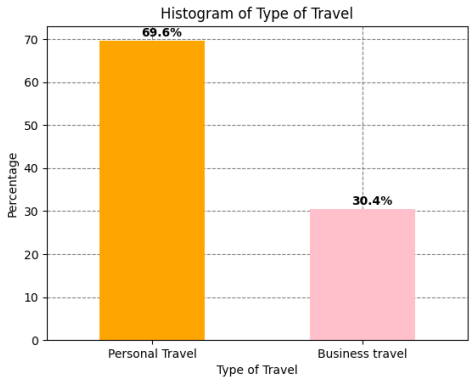


ההסתברויות האפריוריות מראות כי 49.6% נשים ו50.4% גברים. הנתונים מאוזנים לגמרי מתפלגים חצי חצי. הדבר מייצג את המציאות כיוון שהאוכלוסייה בעולם מחולקת שווה גברים ונשים ולכן בסבירות גבוה גם הלקוחות הטסים.

customer type -2



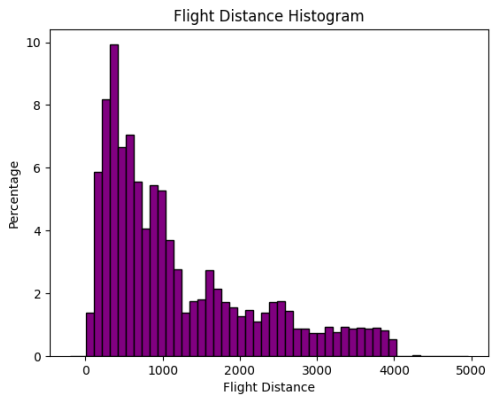
ההסתברויות האפריוריות מראות כי 82% מהלקוחות נאמנים, ו-18% לא נאמנים. הנתונים כלל לא מאוזנים, ניתן לראות שיש הרבה יותר נאמנים. הדבר יכול לייצג את המציאות אם החברה עם שירות טוב, ערך תמורת כסף גבוה, חווית לקוח חיובית (כמו חברת apple לדוגמה).



type of travel -3

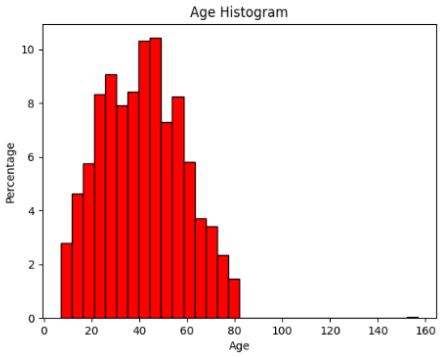
ההסתברויות האפריוריות מראות כי 69.6% בנסיעות אישיות ו-30.4% בנסיעות עסקים. הנתונים לא מאוזנים, הדבר מלמד שיש יותר נסיעות אישיות. לדעתנו הדבר מייצג את המציאות כיוון שרוב האוכלוסייה טסה כדי לנפוש ולא לצורך עבודה.

משתנים רציפים:



flight distance -4

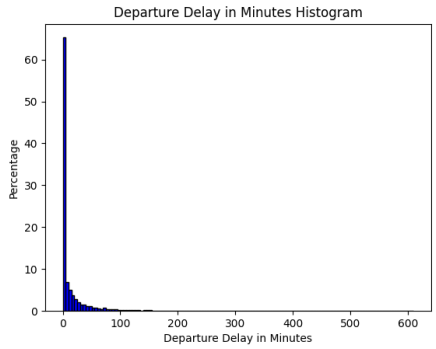
ההיסטוגרמה מחולקת לבינים כאשר כל בין מכיל 100 ק"מ.
ניתן לראות שרוב הטיסות הם עד 1000 ק"מ, כאשר הכי הרבה טיסות (18%) הם בין 200 ל 400 ק"מ.
כל בין בטיסות שמעל 1000 ק"מ מתפלגות יחסית אחיד סביב ה 1.6% . הדבר הגיוני לדעתנו מכיוון שיותר זול לטוס ליעדים קרובים ולכן הביקוש ליעדים אלו גבוה.



age -5

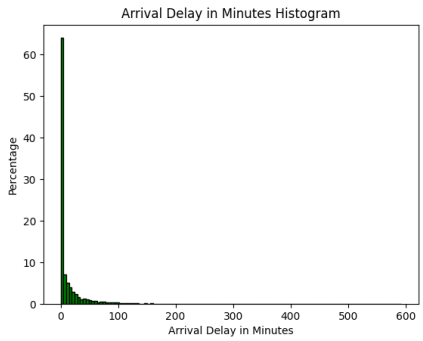
ההיסטוגרמה מחולקת לבינים כאשר בכל בין קיים טווח של 5 שנים. ניתן לראות שהגילאים מתפלגים נורמלית בין גיל 5 עד 85, יש תוצאה חריגה בגיל 160 ואין נתונים על גילאים 0 עד 5.
לדעתנו ההתפלגות מייצגת את המציאות שרוב הגילאים שטסים הם סביב גיל ה 30-50 וככל שאתה צעיר יותר או מבוגר יותר פחות יש לך זמינות לכך.

6- departure delay in minutes



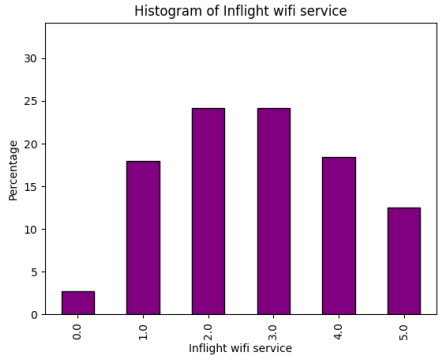
ההיסטוגרמה מחולקת לבינים כאשר בכל בין קיים טווח של 5 דקות איחור. ניתן לראות שהתפלגות האיחורים היא מעריכת, רוב המחולט של הטיסות לא מאחרות כלל ולכן בערך 0-5, הדבר מייצג את המצאות בשגרה.

7- arrival delay in minutes



ההיסטוגרמה מחולקת לבינים כאשר בכל בין קיים טווח של 5 דקות איחור. ניתן לראות שהתפלגות האיחורים היא מעריכת, רוב המחולט של הטיסות מגיעות בזמן, הדבר מייצג את המצאות בשגרה. הגרף כמעט זהה ל (6) מכיוון שאיחור ביציאה גורם באופן ישיר באותו הערך לאיחור בהגעה .

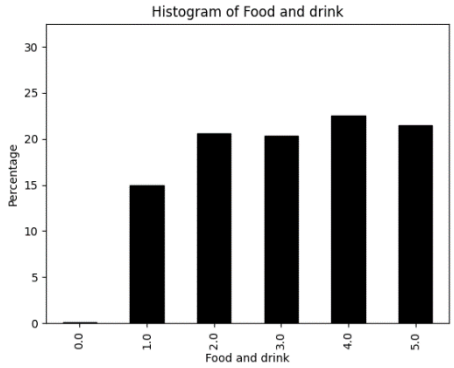
משתנים בדידים שמתייחסים אליהם כרציפים:



8- inflight Wi-Fi service

ניתן ללמוד מההיסטוגרמה שאחוז הטיסות ללא WIFI (ערך 0) נמוך יחסית, וששביעות רצון האנשים מתפלגת נורמלית בין 1 ל 5.

ניראה לנו יחסית מציאותי, לכל אדם יש צורך שונה מהאינטרנט שמשליך כנראה על רמת האינטרנט הרצויה בעיניו.



9- food and drink service

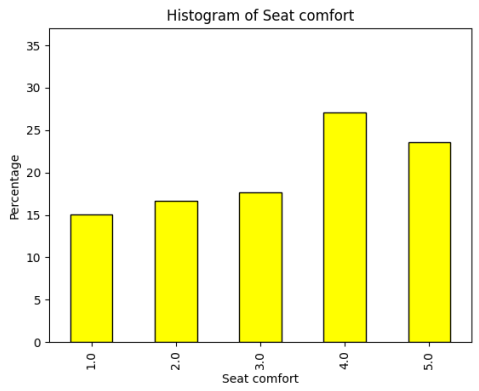
התפלגות שביעות רצון האנשים מהמזון והשתייה מפולגת יחסית אחיד סביב ה 20% לכל דירוג, כאשר 1 נמוך מהאחרים.

לדעתנו הדבר לא מייצג את המציאות בצורה מלאה, מהיכרות אישית וחוויה של הסובבים בדרך כלל המזון והשתייה ברמה נמוכה מהמצופה. ולכן היינו מצפים לראות את רוב התוצאות ב 1,2.

seat comfort -10

ניתן לראות שרוב הביקורות על נוחות המושב הם חיוביות כמעט 50%, בערך 17% מרוצים באופן בינוני והאחרים לא מרוצים.

הדבר לדעתנו מייצג את המציאות כיוון שלרוב בטיסות הכיסאות מרווחים עם התאמות שונות כגון מזגן, כיוון משענת, אוזניות, ידיות לידיים שמעלה את רמת סיפוק הלקוחות.

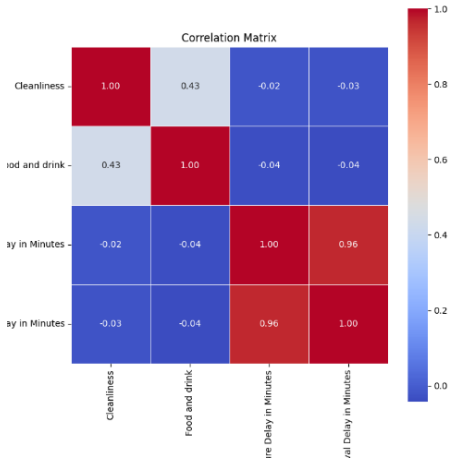


2.3 קשרים בין משתנים- צפויים ולא צפויים :

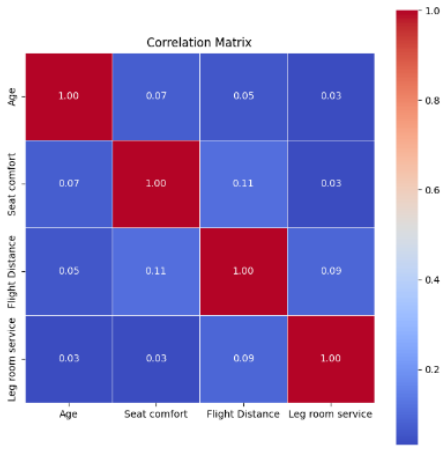
<u>קורלציה בין מאפיינים צפויים:</u>

ניתן לראות בטבלה קורלציה חיובית גבוה מאוד (0.96) בין **departure delay in minutes** לבין **arrival delay in minutes**. הדבר צפוי והגיוני מכיוון שטיסה שמאחרת ביציאה, גם תאחר בשעת הגעה באותו ערך.

בנוסף קיים קורלציה בינונית (0.43) בין **Food and drink** לבין **Cleanliness**, הדבר גם יחסית הגיוני מכיוון שניקיון יכול להשפיעה על התפיסה שלנו לרמת הכנת האוכל וההחזקה שלו, אך היא לא גבוה מכיוון שיש עוד אספקטים לאיכות האוכל כמו הטעם והטריות.



קורלציה בין מאפיינים לא צפויים:



ניתן לראות בטבלה קורלציה נמוכה מאוד (0.07) בין **age** ל **seat comfort**. הדבר לא צפוי, מכיוון שניתן לשאר מידע מקדים כי ככל שאדם יהיה מבוגר יותר יהיה לו פחות נוח בישיבה ממושכת בשל כאבי גב, רגלים. היינו מצפים לקורלציה חלקית לפחות.

בנוסף קיים קורלציה נמוכה (0.09) בין **Flight Distance** ל **Leg room service** . הדבר לא צפוי כיוון שהיינו משארים שככל שהטיסה יותר רחוקה יקדישו לכך מטוסים יותר גדולים שבהם יהיה קיים מקום מרווח יותר לרגליים.

2.4 מאפיינים שנחשוד שבעלי השפעה על משתני המטרה:

לפי בדיקה באינטרנט, חוויות וידע אישי, לדעתנו המאפיינים בעלי ההשפעה הגדולה ביותר על שביעות רצון (פונקציה המטרה) הם:

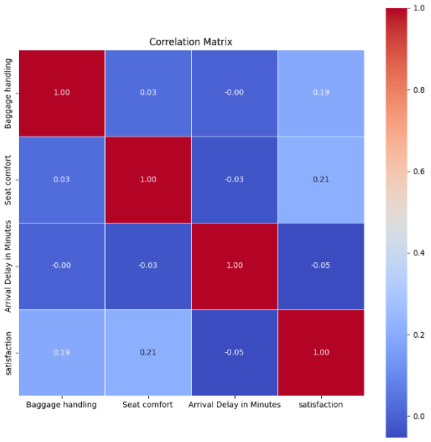
Baggage handling- נוסעים מצפים שהכבודה שלהם תגיע בשלום ובזמן ליעדה. כאשר יש בעיות כמו אובדן, נזק או איחור בהגעת הכבודה, זה יכול לגרום לתחושת חוסר נוחות ואכזבה משמעותית.

Arrival Delay in Minutes- דייקנות הזמנים היא חלק מהותי מציפיות הנוסעים מחברת התעופה מכיוון שמשפיע על לוחות הזמנים של הנוסעים, הגברת הסטרס, פיצוי על עיכובים מה שמצריך טרחה עבור הנוסע.

Seat comfort- מושבים נוחים עם ריפוד איכותי, יכולת להטות את המושב, רוחב מספק ואביזרים משלימים יכולים להפוך את חוויית הטיסה לנעימה יותר, להפחית עייפות וכאבים,

ולשפר את שביעות הרצון הכללית. כאשר חברות תעופה משקיעות בנוחות הישיבה, הן תורמות באופן משמעותי להרגשה הטובה של הנוסעים ולהעדפתם לטוס שוב עם אותה חברה.

2.5 קורלציה בין המשתנים לפונקציית המטרה שחשדנו כבעלי השפעה



אמנם קורלציה של 0.2 היא נמוכה, אבל כאשר הסתכלנו על הטבלת קורלציה בין כל המשתנים ומשתנה המטרה (נספח 99) ראינו כי **Baggage handling ,Seat comfort** הם עם קורלציה גבוה ביחס למשתנים הקיימים.

ראינו גם שיש קורלציה נמוכה מאוד במשתנה Arrival Delay in Minutes **.**

2.6 משתנים שמניחים שניתן להסיר:

לדעתנו המשתנים שכדי להסיר הם **Plane colors**, **Gate location**

אנו כמעט ולא רואים קשר בין משתנים אלו לבין שביעות רצון הלקוח, רוב האנשים לא מיחסים חשיבות לצבעי המטוס או מיקום שער מסוים. בנוסף לכך ההנחה שלנו מתקיימת לפי טבלת הקורלציה (נספח 99) עם מקדם מתאם של 0.01 .

3. איכות הנתונים:

3.1 נתונים חסרים, מה ניתן לומר עליהם

```
dtype: int64
5568    10
5602    16
```

הוצאנו פלט של הרשומות (שורות) שלהם חסר מעל 5 שדות, ראינו שקיימות 2 רשומות כאלה. אחת עם 10 שדות ריקים, ואחד עם 16 שדות ריקים.

Gender	0
Customer Type	1
Age	1
Type of Travel	1
Class	1
Flight Distance	2
Plane colors	0
Inflight wifi service	2
Departure/Arrival time convenient	2
Ease of Online booking	2
Gate location	2
Food and drink	2
Seat comfort	1
On-board service	1
Leg room service	2701
Baggage handling	1
Checkin service	1
Inflight service	1
Cleanliness	1
Departure Delay in Minutes	1
Arrival Delay in Minutes	471
satisfaction	1

הוצאנו פלט של כמות הנתונים החסרים מכל משתנה (עמודה)
ניתן לראות שלרוב המשתנים חסרים נתונים בודדים (0,1,2).
עבור המשתנה A.D.I.M יש 471 נתונים חסרים.
עבור משתנה L.R.S יש 2701 נתונים חסרים.
הדבר מהווה 30% מהנתונים של עמודה זאת.

3.2 נתונים שאינם הגיוניים

לאחר בדיקה של הנתונים לכל משתנה בנפרד ראינו את התוצאות החריגות האלו:

Age- ראינו 2 רשומות עם ערכי גיל גדולים מ 120 (156,157) הבנו שהם לא הגיוניים.

Flight Distance- ראינו רשומה אחת חריגה של מרחק שלילי (204-), ערך מרחק שלילי לא אפשרי.

Gate location- ראינו רשומה אחת חריגה של דירוג מיקום (999), ערך לא מהאופציות לכן לא אפשרי.

Inflight service- ראינו רשומה אחת חריגה של דירוג שירות הטיסה (0), מכיוון שרשומה זאת יחידה הסקנו שהיא לא מהאופציות ולכן לא אפשרית.

class- ראינו רשומה אחת עם מלל לא מהאופציות של המשתנה, בנוסף ראינו 1311 רשומות עם הערך Unknown שהוא גם לא באופציות.

תובנות נוספות:

Inflight Wi-Fi service - קיימות רשומות עם הערך 0 במשתנה זה, הנחנו כי ערך זה מציין שאין WIFI זמין בטיסה.

Food and drink- קיימות רשומות עם הערך 0 במשתנה זה, הנחנו כי ערך זה מציין שאין חלוקת שתייה ואוכל בטיסה.

הכנת נתונים:

4. על פי הצורך, בצעו ונמקו בחירת מאפיינים שביצעתם:

4.1 השמטת תצפיות בעלי חוסר רב:

בהתאם לסעיף 3.1 הנתונים החסרים מכל רשומה (שורה).

בחרנו למחוק את 2 הרשומות בעלי כמות שדות ריקים הגדולה מ 5, הדבר טוב מכיוון שבכך אנחנו לא צריכים למלא שדות ריקים ובכך להגדיל את הרעש וההטיה.

הדבר יכול לפגוע בכך שנקטין את גודל סט האימון, מכיוון שכמות נתונים גדולה יותר מייצגת יותר את המציאות. אבל במקרה זה 2 רשומות מתוך 9000 זה כמות זניחה.

לאחר הסרת 2 הרשומות קיבלנו שהנתונים החסרים הם במשתנים:

Gender	0
Customer Type	0
Age	0
Type of Travel	0
Class	0
Flight Distance	0
Plane colors	0
Inflight wifi service	0
Departure/Arrival time convenient	0
Ease of Online booking	0
Gate location	0
Food and drink	0
Seat comfort	0
On-board service	0
Leg room service	2700
Baggage handling	0
Checkin service	0
Inflight service	0
Cleanliness	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	470
satisfaction	0

משתנה A.D.I.M יש 470 נתונים חסרים,

בנוסף לכך משתנה זה מושפע בצורה ישירה ממשתנה D.D.I.M בעצם טיסה שמאחרת ביציאה גם תאחר בשעת הגעה באותו ערך קורלציה כמעט 1 לפי טבלת 99 (בנספחים) חשש למולטיקולינריות,

בחרנו לוותר על משתנה זה.

משתנה L.R.S יש 2700 נתונים חסרים.

הדבר מהווה 30% מהנתונים, השלמה של כמות נתונים זאת לדעתנו תפגע בהימנות ונכונות הנתונים ולכן נבחר למחוק את המשתנה.

4.2 השלמה מושכלת של נתונים ערכים חריגים במידה ואפשר:

בהתאם לסעיף 3.2 הנתונים בחלק מהמשתנים היו חריגים, לכן:

Age- בחרנו לשנות את ערכים אלו בערך הממוצע של המשתנה.

Flight Distance- נשנה את הערך במרחק טיסה ממוצע.

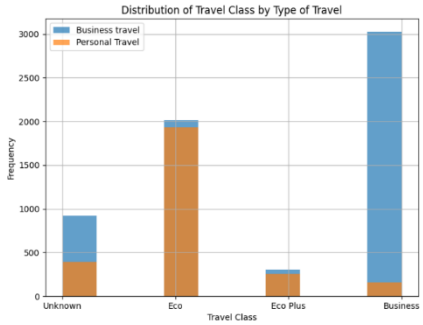
במקרה זה מדובר על רשומות חריגות בודדות מילוי ערכים חסרים בערך הממוצע הוא גישה פשוטה ויעילה לשמירה על כמות הנתונים ועל האיזון הכולל במערך הנתונים.

Gate location- נשנה לערך הכי נפוץ .

Inflight service- נשנה לערך הכי נפוץ .

גם במקרה זה מדובר על רשומה בודדת לכל משתנה ולכן שינוי הערכים החריגים בנתונים הכי נפוצים מאותו משתנה לא תגרום להטיה, כאן לא נשתמש בממוצע מכיוון שהמשתנים קטגוריאליים ולא נוכל לקבל ערך שהוא לא מספר שלם.

Class- החלטנו לבדוק את היחס בין המשתנים Class, Type of travel כי מבחינה לוגית הגיונית היה ניראה לנו שקיים קשר. יצרנו היסטוגרמה המראה את התדירות של סוג הטיול מול מחלקת הטיסה.



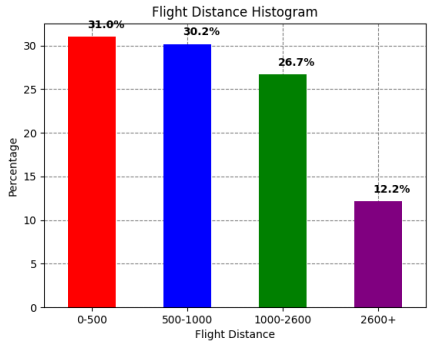
ניתן לראות כי כאשר המשתנה של נוסע הוא מסוג טיול עסקי (business travel) כמעט כל הערכים שלו מהמשתנה מסוג מחלקת טיסה (class) היו business.
בהתאמה כאשר המשתנה של נוסע הוא מסוג טיול אישי (personal travel) כמעט כל הערכים שלו מהמשתנה מסוג מחלקת טיסה (class) היו Eco.

לכן החלטנו שאת השדות unknown ו "IT IS SO BORING…" מה שלא מהאופציות במשתנה travel class, נחליף ל business ו Eco בהתאם לסוג השדה ב type of travel.

במילים פשוטות : Type of travel = Business travel --> Class = Business
Type of travel = Personal Travel --> Class = Eco

5. על פי הצורך, תנו טיפול פרטני במאפיינים:

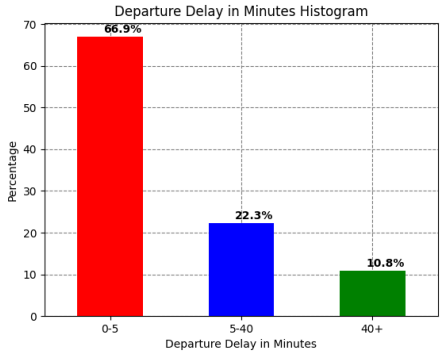
5.1 דיסקרטיזציה של משתנה רציף flight distance



חלקנו את הטווחים ל 4 סיפים כיוון שראינו מדרגות בהיסטוגרמה. טווח של 0-500, 500-1000, 1000-2600, 2600+

יצרנו עמודה חדשה בשם **Flight Distance Rank** עם ערכים 1,2,3,4 כאשר 1 הטווח הכי קצר.

5.2 דיסקרטיזציה של משתנה רציף departure delay in minutes



חלקנו את הטווחים ל 3 סיפים כיוון שראינו שרובם של האיחורים נופל בין 0-5, חלקם הקטן בין 5 ל 40 וכמעט ולא קיים בכלל גדול מ 40.

יצרנו עמודה חדשה בשם **Departure Delay Rank** עם ערכים 1,2,3 כאשר 1 האיחור הכי קטן.

5.3 גזירת מאפיינים חדשים מהמשתנים On-board service ו inflight service

לא הבנו את הסיבה לשני משתנים המייצגים דבר דומה של איכות השירות ולכן יצרנו משתנה חדש בשם **service quality** המייצג את הממוצע בין Leg room service ו Seat comfort.

5.4 גזירת מאפיינים חדשים מהמשתנים Baggage handling ו Check-in service

חשבנו כי יש אולי קשר בין המשתנים כיוון שחלק מתהליך הצק קיים טיפול בכבודה.

יצרנו משתנה חדש בשם **Baggage service** המייצג את הממוצע בין 2 המשתנים.

נספחים:

טבלת קורלציה (99) בין משתנה המטרה לכל המשתנים המספריים: